Alternative splicing: a bioinformatics perspective

Elisa Napolitano Ferreira,[a] Pedro A. F. Galante,[b] Dirce Maria Carraro[a] and Sandro José de Souza*[b]

DOI: 10.1039/b702485c

The degree of diversity at the transcriptome and proteome levels generated by alternative splicing is astonishing. In this review, we discuss several issues related to alternative splicing with a special emphasis on identification strategies based on bioinformatics.

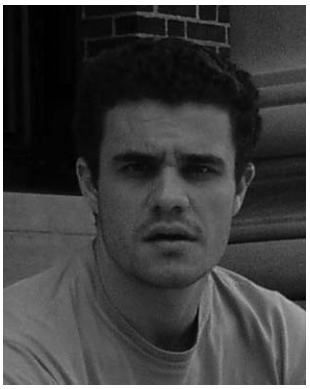
Introduction

Splicing refers to the processing of pre-mRNA that joins the exons and removes the intervening sequences (introns). Although the chemical nature of the splicing reaction is very simple, based on two trans-esterification reactions, its execution by cells in a biological context is extremely complex. A cellular particle, named spliceosome, composed of five small nuclear RNAs and more than 200 proteins, is responsible for splicing in the nucleus of basically all eukaryotic cells. Immediately after the discovery of introns, Gilbert[1] predicted the existence of variants produced by the alternative choice of new exon/intron borders. This process, that we now call alternative splicing (AS), was soon after documented[2,3] and believed to be a rare phenomenon occurring in approximately 5% of higher eukaryotes genes.[4] Currently estimates suggest that more than 60% of all human genes code for two or more splicing variants.[5] Similar estimates have been made for mice.[5] This phenomenon in part explains the paradox between the small number of genes compared to the high complexity of higher eukaryotes and is a major contributor to proteomic

[a]*Ludwig Institute for Cancer Research, São Paulo branch, Hospital A. C. Camargo, São Paulo, Brazil. E-mail: eferreira@ludwig.org.br; dcarraro@ludwig.org.br*
[b]*Ludwig Institute for Cancer Research, São Paulo branch, Hospital Alemao Oswaldo Cruz, São Paulo, Brazil. E-mail: pgalante@ludwig.org.br; sandro@ludwig.org.br*

Elisa Napolitano Ferreira

Elisa N. Ferreira is a PhD student from the University of São Paulo—Brazil interested in the molecular diversity of the human transcriptome generated by alternative splicing.



Pedro A. F. Galante

Pedro A. F. Galante from the Ludwig Institute for Cancer Research, São Paulo branch is a young bioinformatician with significant contributions in genomics and bioinformatics.

Dirce Maria Carraro

Dirce Carraro is a senior scientist at the Hospital AC Camargo in São Paulo—Brazil. Her major interests are related to cancer genomics.

Sandro José de Souza

Sandro José de Souza is an Associated Member and head of the Laboratory of Computational Biology at the Ludwig Institute for Cancer Research–São Paulo branch. Dr de Souza holds a PhD in biochemistry and has made many contributions in molecular evolution, genomics and bioinformatics.

diversity.[6] This diversity at the protein level is achieved because most AS occurs within the coding region.[7] These AS variations can be classified in three major categories. The most frequent one is exon skipping (42% in human), in which an exon (or a group of continuous exons) can be alternatively used in different messages.[5] Less frequent events include the alternative usage of donor (23% of all events in human) and acceptor sites (26% in human) and intron retention (09% in human).[5] Fig. 1 shows a schematic view of these different types of AS.

In the last few years a correlation between AS and different features of human diseases has become a central issue in biomedical research.[8] Particularly interesting is the association between splicing variants and cancer.[9] Several cancer-associated variants have been identified and they might contribute to the development of diagnosis and prognosis protocols and also serve as therapeutic targets.[10] Therefore, obtaining a complete catalogue of all human transcripts and their involvement in several biological processes would allow a better comprehension of development mechanisms, tissue-specific processes and also enable characterization, comprehension and treatment of diseases.

Central to the studies linking AS to disease is how cells regulate constitutive and alternative splicing (by constitutive splicing, we mean the excision of introns having always the same borders, *ie*, absence of AS). The emerging scenario couldn't be more complex. There are three basic sequence features involved in splicing regulation: i, the exon/intron junctions at both the 5′ and 3′ ends of the introns (donor and acceptor sites, respectively); ii, the branch site located upstream of the acceptor site and iii- the polypyrimidine tract located between the acceptor site and the branch site. It is quite clear that these elements *per se* cannot regulate constitutive and alternative splicing since they are short and not very informative. Bioinformatics and experimental approaches have unraveled a large number of sequence elements that contribute to the regulation of splicing. Roughly, there are two additional types of *cis*-acting elements, the enhancers and the silencers with different and obvious effects on splicing. The nomenclature of these elements takes into consideration not only their effect on splicing but also their location in the gene, either exonic or intronic. They are called exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and the intronic splicing silencers (ISSs). These elements are present in basically all genes in complex eukaryotes. Fig. 1 also illustrates the relative position of all these elements in a hypothetical gene. Most of the enhancers act as binding sites for a class of RNA-binding proteins called SR (serine/threonine) while most of silencers are binding sites for hnRNPs (heteronuclear ribonucleoproteins), with many exceptions being reported.[11]

We know today that AS is tightly linked to other RNA processing phenomena inside the cell,[12,13] transcription is a clear example. It has been shown, for example, that the pausing pattern of RNA polymerase II along the gene being transcribed can affect the selection of exon/intron borders in the respective message.[13] AS is also associated with the phenomenon of nonsense-mediated decay (NMD) responsible for the degradation of messages presenting a premature stop codon.[14]

Ways to detect alternative splicing

Bioinformatics has been fundamental in the characterization of many new splicing variants. While the *ab initio*

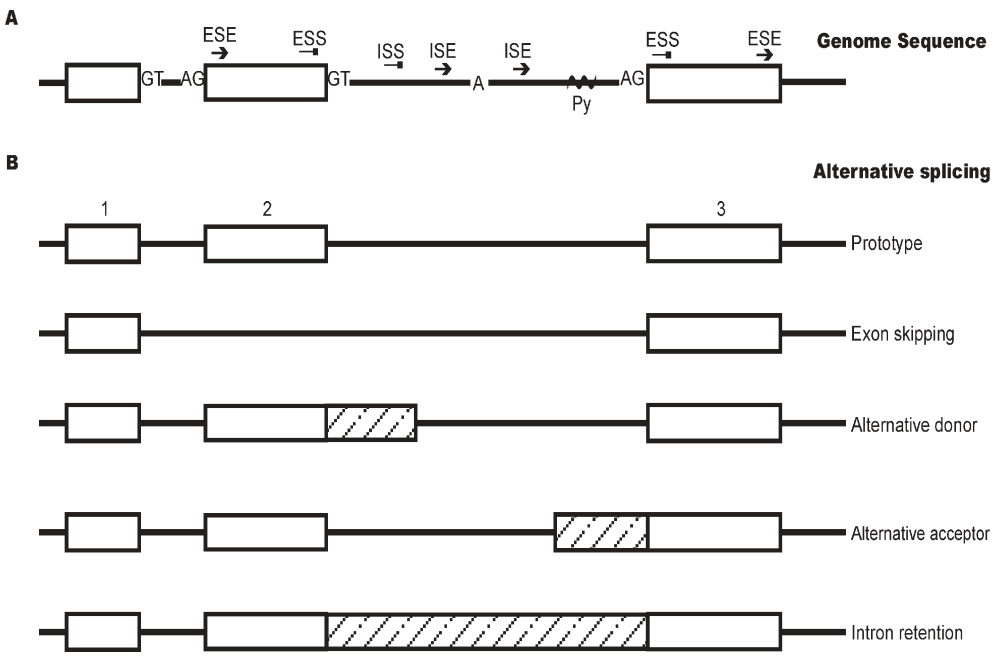


Fig. 1 Schematic view of constitutive and alternative splicing for a hypothetical gene with the major *cis*-acting elements represented. A represents the genome sequence for a hypothetical gene. B represents all major types of AS. Exon 1 is subjected to constitutive splicing. Exons 2 and 3 are subjected to alternative splicing. The major *cis* regulatory elements are also shown. GT: donor site; AG: acceptor site; Py: polypyrimidine tract; A: branch site; ESE: exonic splicing enhancer; ESS: exonic splicing silencer; ISE: intronic splicing enhancer; ISS: intronic splicing silencer.

identification of splicing variants has been hampered by the lack of information on basic aspects of splicing and its intrinsic complexity, the processing of experimental data has been speeded up by the development of computational tools and strategies. Below we describe the major ways to detect splicing variants giving a special emphasis on their computational aspects.

cDNA sequences

The large number of expressed sequences identified and available at public databases in the last decades became a potential source for large-scale identification of splicing variants. EST collections are especially suitable once they are originated by partial sequencing of mRNAs from different tissues, developmental and pathological stages.

One of the simpler approaches consists in aligning EST sequences against each other or against mRNA sequences to search for regions of insertion or deletion.[15] A significant improvement was achieved by the alignment of both EST and mRNA sequences to the genome sequence, which enables the definition of exon/intron boundaries.[16] In this sense the development of algorithms, like Sim4,[17] that more precisely define the exon/intron borders by adjusting the alignments to force the use of canonical donor and acceptor sites and contributed to more precise inferences.

Currently the most widely used strategy consists of two steps. First, the alignment of cDNA sequences against the human genome followed by the comparison of exon/intron borders among all expressed sequences of the same gene.[10,18] This allows the evaluation of a broader number of sequences rather than the traditional pairwise alignment. To date, the information generated by many cDNA-genomic alignment projects are available in different AS databases, which are listed in Table I.

Another valuable strategy for AS identification is the cross-species alignment of ESTs against genome sequences, especially when performed between closely related organisms. The alignment of mouse transcript sequences to the human genome identified 8921 AS events not represented in human transcript sequences, which was considered a prediction of novel human variants.[24] Another study comparing human, mouse and rat datasets predicted 320 novel alternative human exons based on transcript sequences of mouse and rat. RT_PCR experiments validated more than 50% of the tested exons.[25] Although this approach has been useful for the identification of new splicing variants, there are several limitations. The most serious is the differential representation of tissues and cells from different organisms in the public databases.

Although there are ~7.9 million human ESTs in the respective database (dbEST release 011907) they do not cover the entire human transcriptome reducing the repertoire of detectable AS isoforms. This can be illustrated by the shorter number of ESTs from normal tissue in comparison to tumor counterpart affecting studies focused in identification of tumor-associated variants.[26] Other problems have been highlighted when aligning EST to genome sequence including poor quality alignments, vector contamination, genomic DNA contamination and incomplete splicing that compromise the identification of bona fide splicing isoforms, although many of these problems can be solved through bioinformatics filters.

A serious limitation on the utilization of ESTs for the identification of splicing variants is their fragmentary nature. Since ESTs are in average 500 bp long, AS events are identified in a individual basis without the context of the whole transcript. The precise identification of all AS events in a single transcript is achieved by the sequencing of complete cDNA clones as done by several initiatives including MGC (Mammalian Gene Collection).[27]

Microarray AS platform

Microarray technology is a powerful approach for large-scale expression analysis of genes in a large amount of samples. Therefore it is suitable for the characterization of expression profile of the different transcripts of a gene. Many platforms have been designed to specifically measure the abundance of splicing variants.[28] Most of these approaches depend on the design of specific probes spanning alternative exon junctions. Although a previous knowledge of exon boundaries is necessary, this approach is useful for evaluating splicing profiles in different tissues. Johnson *et al.*[29] analyzed 10 000 human cDNAs through an exon junction platform hybridized with 52 tissues and could determine tissue distributions for thousands of known and novel alternative splicing events. In another study Gardina *et al.*[30] constructed a platform with 1.4 million probe sets representing one million exons based on mRNA sequences and *ab initio* computational predictions. This platform was hybridized with 10 matched pairs of normal and colon primary tumor and resulted in 189 new putative splicing events, of which 43 were submitted to RT_PCR with a validation rate of approximately 30%.

Another valuable platform is the tilling array, in which probes are designed across all loci in a chromosome or whole genome. Depending on the design of the array, however, the precise definition of the exon/intron border is somehow limited and the platform is better suitable for the detection of exon skipping and intron retention events. Furthermore, not all exons can be identified and the definition of the expression level of exons bearing repetitive elements, such as Alu sequences, is seriously limited because of cross-hybridization (in some platforms such repetitive elements are avoided in the stage of probe design).

The microarray technology, as with most of the other technologies, does not allow the unambiguous identification of the exon/intron organization of a full transcript since it is not possible to characterize how most of the exons are connected.

cDNA libraries enriched with splicing variants

Another interesting approach for the identification of splicing variants is the construction of AS-enriched cDNA libraries. This approach is based on the availability of two full-length cDNA libraries constructed from distinct samples and results in a third library composed by clones representing AS events between the two parental libraries. The enrichment step is based on the assumption that heteroduplexes formed between two different splicing variants

Table 1 Examples of AS databases

DB	Release	Source	# of Entries	URL
ASDB[19]	2.1	Swiss-Prot mRNAs from GenBank	1762	**cbcg.nersc.gov/asdb**
ASAP[20]	Jan06	ESTs from UniGene mRNA from GenBank Genome from NCBI	89 078	**www.bioinformatics.ucla.edu/ASAP2**
ASD[21]	3	ESTs from dbEST mRNAs from ensembl	65 451	**www.ebi.ac.uk/asd**
EASED[22]	#131	ESTs from dbEST mRNA from ensembl	33 100	**eased.bioinf.mdc-berlin.de**
Hollywood[23]	Jan06	ESTs and mRNAs from GenBank	37 366 alternative exons	**hollywood.mit.edu**

from the same gene can be retrieved by using either biotinilated random oligos[31] or single-strand binding proteins[32] This method is a very powerful approach since any kind of alternative splicing event may be identified and a previous knowledge of the transcripts of both parental libraries is unnecessary. Although promising, this is a laborious approach depending on extensive cDNA librariy construction and sequencing. Furthermore, at least in their present form, such methods do not allow the complete characterization of all splicing borders in a single transcript.

AS in the new genomics era

The technological advances in DNA sequencing in the last few years promise to revolutionize the biomedical sciences.[33] Based on the progress made with the new sequencers already in the market (see http://www.454.com and http://www.solexa.com) and others that will appear soon (from Applied Biosystem and Helicos) it is reasonable to estimate that in a few years we will be able to sequence a human genome (~3GB of DNA) for less than US$2000. This will generate an amount of data that even for today's standards is astonishing. Regarding AS we will be able to: (i) generate a complete spatial/temporal map of all splicing variants in human cells, (ii) sequence both expressed and genome sequence for the same subjects to associate germline and somatic mutations with the expression pattern of splicing variants and (iii) associate AS patterns with specific features of many biological situations, including pathologies. A critical issue with these technologies is their read length, usually shorter than the conventional Sanger-based sequencing. Based on that, a whole transcriptome shotgun approach will probably cover most of the alternative exon/intron sites but not the combination of sites in a single transcript.

In terms of systems biology, one will be able to perturb the expression/function of components of a given system and exhaustively evaluate the effect of those perturbations in the expression of all genes in the system. In principle, we should have available a map of all splicing variants from cells in which components of the spliceosome were perturbed (through either knockout or over-expression experiments). A comparison of the AS pattern among these cells will certainly help to elucidate the function of many splicing factors.

The challenges for bioinformatics are hard to measure. First, we will need to integrate all these types of data in a concise and productive way. There will be a need for more sophisticated mining tools to manage the enormous amount of data In these aspects, the databases listed in Tables 1 will have a pivotal role in making all these data accessible in a friendly way to regular biologists. Finally, we should be able to more effectively integrated data derived from many of the "omics" approaches into strategies aiming to perform dynamic modeling of biological systems. This may be important as a source of hypotheses that can then be tested in an experimental model.

Final comments

The experimental and computational approaches described above have been and will continue to be important for a better understanding of both constitutive and alternative splicing. Despite the great progress obtained in the last decade we are still quite far from having a complete understanding of the phenomena in most of the biological contexts.

It is our feeling that we are just starting to explore the enormous complexity generated by AS in the transcriptome and proteome. It is important that we continue to generate data from large-scale approaches, especially now in the light of these new sequencing technologies. For instance, no single cDNA library has been exhaustively sequenced to determine all the splicing variants in a cell/tissue. In parallel, this has to be coupled with strategies that will allow a better understanding of the mechanisms responsible for the regulation of AS.

References

1 W. Gilbert, *Nature*, 1978, **271**, 501.
2 P. Early, J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall and L. Hood, *Cell*, 1980, **20**, 313–318.
3 M. G. Rosenfeld, C. R. Lin, S. G. Amara, L. Stolarsky, B. A. Roos, E. S. Ong and R. M. Evans, *Proc. Natl. Acad. Sci. U. S. A.*, 1982, **79**, 1717–1721.
4 P. A. Sharp, *Cell*, 1994, **77**, 805–815.
5 E. Kim, A. Magen and G. Ast, *Nucleic Acids Res.*, 2007, **35**, 125–131.
6 B. R. Graveley, *Trends Genet.*, 2001, **17**, 100–107.
7 M. Zavolan and E. van Nimwegen, *Curr. Opin. Struct. Biol.*, 2006, **16**, 362–367.
8 E. Buratti, M. Baralle and F. E. Baralle, *Nucleic Acids Res.*, 2006, **34**, 3494–3510.
9 O. L. Caballero, S. J. de Souza, R. R. Brentani and A. J. Simpson, *Dis. Markers*, 2001, **17**, 67–75; B. M. N. Brinkman, *Clin. Biochem.*, 2004, **37**, 584–594; J. P. Venables, *Cancer Res.*, 2004, **64**, 7647–54.
10 N. Kirschbaum-Slager, R. B. Parmigiani, A. A. Camargo and S. J. de Souza, *Physiol. Genomics*, 2005, **21**, 423–432.
11 F. Pagani and F. E. Baralle, *Nat. Rev. Genet.*, 2004, **5**, 389–396.
12 K. M. Neugebauer, *J. Cell Sci.*, 2002, **115**, 3865–3871.
13 A. R. Kornblihtt, *Nat. Struct. Mol. Biol.*, 2006, **13**, 5–7.
14 F. Lejeune and L. E. Maquat, *Curr. Opin. Cell Biol.*, 2005, **17**, 309–315.

15 J. Burke, H. Wang, W. Hide and D. B. Davison, *Genome Res.*, **8**, 276–290; D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich and P. Bork, *FEBS Lett.*, 2000, **474**, 83–86.
16 B. Modrek and C. J. Lee, *Nat. Genet.*, 2003, **34**, 177–180.
17 L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin and W. Miller, *Genome Res.*, 1998, **9**, 967–974.
18 P. A. Galante, N. J. Sakabe, N. Kirschbaum-Slager and S. J. de Souza, *RNA*, 2004, **10**, 757–765.
19 I. Dralyuk, M. Brudno, M. S. Gelfand, M. Zorn and I. Dubchak, *Nucleic Acids Res.*, 2000, **28**, 296–297.
20 N. Kim, A. V. Alekseyenko, M. Roy and C. Lee, *Nucleic Acids Res.*, 2006, **00**, D1–D6.
21 S. Stamm, J. J. Riethoven, V. L. Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N. L. Barbosa-Morais and T. A. Thanaraj, *Nucleic Acids Res.*, 2006, **34**, D46–D55.
22 H. Pospisil, A. Herrmann, R. Bortfeldt and J. Reich, *Nucleic Acids Res.*, 2004, **32**, D70–D74.
23 D. Holste, G. Huo, V. Tung and C. B. Burge, *Nucleic Acids Res.*, 2006, **34**, D56–D62.
24 Z. Kan, P. W. Garrett-Engele, J. M. Johnson and J. C. Castle, *Nucleic Acids Res.*, 2005, **33**, 5659–66.
25 F. C. Chen, C. J. Chen, J. Y. Ho and T. J. Chuang, *BMC Bioinformatics*, 2006, **7**, 136.
26 M. Roy, Q. Xu and C. Lee, *Nucleic Acids Res.*, 2005, **33**, 5026–5033.
27 R. L. Strausberg, E. A. Feingold, L. H. Grouse, J. G. Derge, R. D. Klausner, F. S. Collins, L. Wagner, C. M. Shenmen, G. D. Schuler, S. F. Altschul, B. Zeeberg, K. H. Buetow, C. F. Schaefer, N. K. Bhat, R. F. Hopkins, H. Jordan, T. Moore, S. I. Max, J. Wang, F. Hsieh, L. Diatchenko, K. Marusina, A. A. Farmer, G. M. Rubin, L. Hong, M. Stapleton, M. B. Soares, M. F. Bonaldo, T. L. Casavant, T. E. Scheetz, M. J. Brownstein, T. B. Usdin, S. Toshiyuki, P. Carninci, C. Prange, S. S. Raha, N. A. Loquellano, G. J. Peters, R. D. Abramson, S. J. Mullahy, S. A. Bosak, P. J. McEwan, K. J. McKernan, J. A. Malek, P. H. Gunaratne, S. Richards, K. C. Worley, S. Hale, A. M. Garcia, L. J. Gay, S. W. Hulyk, D. K. Villalon, D. M. Muzny, E. J. Sodergren, X. Lu, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madan, A. C. Young, Y. Shevchenko, G. G. Bouffard, R. W. Blakesley, J. W. Touchman, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. I. Krzywinski, U. Skalska, D. E. Smailus, A. Schnerch, J. E. Schein, S. J. Jones and M. A. Marra, Mammalian Gene Collection Program Team, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 16899–16903.
28 M. Cuperlovic-Culf, N. Belacel, A. S. Culf and R. J. Ouellette, *OMICS*, 2006, **3**, 344–357.
29 J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton and D. D. Shoemaker, *Science*, 2003, **302**, 2141–2144.
30 P. J. Gardina, T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams and Y. Turpaz, *BMC Genomics*, 2006, **7**, 325.
31 A. Watahiki, K. Waki, N. Hayatsu, T. Shiraki, S. Kondo, M. Nakamura, D. Sasaki, T. Arakawa, J. Kawai, M. Harbers, Y. Hayashizaki and P. Carninci, *Nat. Methods*, 2004, **3**, 233–239, Epub 2004 Nov 18.
32 G. Thill, V. Casteli, S. Pallud, M. Salanoubat, P. Wincker, P. de la Grange, D. Auboet, V. Schachter and J. Weissenbach, *Genome Res.*, 2006, **16**, 776–786.
33 E. R. Mardis, *Genome Biol.*, 2006, **7**, 112.